# High Performance Computing Facilities for the Next Millennium

## Mass Storage

SC99 Tutorial
November 14,1999

*Keith Fitzgerald*
Mass Storage, Group Leader
Kfitz@nersc.gov

# Outline

- Selecting (yet another)  New Storage System
- Configuring and Tuning HPSS
- Bandwidth VS. Capacity
- Storage Resource Management
- Data Intensive Computing
- “PROBE”  (... an HPSS Test Environment)
- NERSC Storage Capabilities (if time)

# Selecting a Storage System

- Storage Solutions can be categorized as:
  - Backup systems (ADSM, Networker, etc)
  - Data Migration systems (Cray DMF, SAMFS, etc)
  - Hierarchical Storage Management systems (HPSS, UniTree, CFS, etc)
- Backup solutions may be appropriate if:
  - All data can fit concurrently on the compute server's disks.
- DMF solutions can be effective when
  - You don't mind using supercomputer cycles to provide storage capabilities and your compute server offers the solution and you plan to upgrade with a compatible compute server....
  - OR you can afford to retain the compute server as a file server
- Otherwise, you're probably a candidate for an HSM!
  - Most Supercomputer Centers fall into this category
  - A disadvantage is that users must store and retrieve data

# NERSC Storage Requirements

- Vendor support
- Scalability (should be able to enhance individual capabilities without replacing a monolithic storage server)
- High aggregate transfer rate (many concurrent requests)
- High transfer rates on individual files
- Reliability and Stability
- Support for very large files
- Support for high performance (and capacity) devices
- Accommodate a large number of files (10**7+)
- Good management interface

# NERSC Selected HPSS

- High Performance Storage System (HPSS) is an IBM service offering which was developed and is being enhanced through a collaboration between IBM and five national labs (LANL, LLNL, ORNL, Sandia, LBNL).

  *Note: NERSC did a zero based evaluation*
- HPSS is a distributed HSM based upon the IEEE mass storage reference model.
- HPSS supports a “3rd party” data transfers
- HPSS is designed to:
  - Handle very large file sizes
  - Accommodate a large number of files
  - Can handle a large number of servers and storage devices
  - “Stripes” storage devices and network interfaces to achieve high transfer rates on individual files.
  - Interfaces to DFS via a “DMIG” interface
  - see http://www.sdsc.edu/hpss for more information

# Configuring a Storage System

## Always based upon *user requirements!!!*

- An HSM generally contains the following elements:
  - An archive (robotic storage and perhaps a shelf operation)
  - Archival storage devices (generally tape drives)
  - Optional intermediate cache devices (generally disks)
  - Storage server(s)
  - Network interface(s)

- Hardware configuration is based upon
  - YOUR STORAGE BUDGET!
  - Robotic archive size determined by  online data requirement

# Cont' Configuring a Storage System

## Always based upon *user requirements!!!*

- Number of archival devices is a complex calculation (guess) based aggregate bandwidth requirements and data access patterns.
- Intermediate cache determined by file sizes and data access patterns.
- Storage servers determined by bandwidth requirements, number and speed of devices, capability of the servers, and data access patterns.
- Network interfaces based upon bandwidth requirements, number of servers, capability of each server, and capability of the interface.

# HPSS Configuration

- HPSS configuration is based upon “storage classes” (SC). Each SC has an associated migration and purge policy.

  (e.g.. SC1=IBM-SSA-disk, SC2=maxstrat-disk, SC3=STK-9840)

- The various SC’s can then be grouped into storage hierarchies (up to five levels deep .... )

  (e.g.. SH1=SC1->SC3; SH2=SC2->SC3; SH3=SC1->SC2->SC3)

- The concept is further refined by configuring storage hierarchies into “classes of service” (COS).

  (e.g.. COS1=SH1, COS2=SH2, COS3=SH3)

# Cont' HPSS Configuration

- Users can select the COS .. or it will default based upon file size
- Physical devices are associated with "movers"
- Tape devices are shared (not tied to a particular storage class)
- Physical media is associated with a storage class
- Physical media may be accessed in parallel (striped) as a single virtual volume. This increases speed at the cost of reliability.

# NERSC HPSS Configuration

## (This is a VERY simple configuration!)

- Storage classes (SC)
  - SC1 3590-tape 1MB virtual volume archive
  - SC2 maxstrat-disk 16MB virtual volume large file cache
  - SC3 SSA disk 16MB virtual volume medium file cache
  - SC4 SSA disk 1MB virtual volume small file cache
- Storage Hierarchies (SH)
  - SH1 SC2->SC1 (large-file-cache to archive)
  - SH3 SC3->SC1 (medium-file-cache to archive)
  - SH4 SC4->SC1 (small-file-cache to archive)

# Cont’ NERSC HPSS Configuration
## (This is a VERY simple configuration!)

- Classes of Service (COS)
  - COS1 SH1 large files 100MB- 32GB
  - COS2 SH3 medium files 2MB-100MB
  - COS1 SH4 small files 0MB- 2MB

(We plan to add dual copy and striped classes of service ASAP)

(We also have a second HPSS system similarly configured)

# Tuning the Storage System
## (A never-ending job ......)

- Tune your "hardware configuration" .....
  - Supply sufficient network bandwidth
  - Enough intermediate cache? (test cache hit rates and residency)
  - Enough individual cache devices? (large disks thrash)
  - Enough CPU and memory power to operate everything efficiently?
  - Enough archival storage capacity for projected capacity demands?
- Tune the actual storage server network interfaces
- Tune the storage server operating systems (swap, networking, etc)
- Tune HPSS ....
  - Setup network options based upon clients and subnets
  - Assign hosts to specific interfaces
  - Optimize server-to-server buffer sizes
- MONITOR usage patterns and response times .....

# Storage Resource Management

- Possible solutions:
  - Anarchy ... at the mercy of any valid user
  - Classic quota mechanism ... space based and not supported in HPSS
  - Charging ....
- A classic storage quota mechanism is not an effective mechanism for archival storage management..... A single large user can saturate your entire network, CPU, memory, cache, and archival devices without exceeding a space quota by just replacing files.
- We wanted to:
  - Account for (and justify) our storage resources
  - Apply storage resources based upon programmatic priorities
  - Constructively influence user behavior by developing a scheme that emphasizes factors that actually use critical resources! (There’s more to storage resources than just a directory entry and some space on a tape. It’s touching the data that hurts!)

# NERSC Storage Resource Units

- What we really implemented is a form of charging which we CALL a “storage quota” because it has a finite value which is allocated as part of our normal “ERCAP” resource request process (a yearly storage allocation).
- We measure usage in terms of Storage Resource Units (SRU)
- A user’s SRU allocation is decremented monthly based upon: Bytes transferred, files in storage, and amount of data in storage.
- The actual formula is:

$$\begin{aligned} \text{SRU} = \; & 4.0 \quad (\text{Gigabytes-of-I/O}) \; + \\ & .4 \quad (\text{Gigabytes-in-storage}) \; + \\ & 0.0012 \; (\text{Number-of-files}) \end{aligned}$$

- Users are notified of their usage monthly via email (& web)
- Abnormal (over quota) behavior is projected and users advised.
- Exceeding your “storage quota” never automatically results in denial of storage service!

# Bandwidth VS. Capacity
## (How many SRU's can we provide?)

- The following guidelines are used in setting our yearly storage allocations ....
- We believe that I/O bandwidth (rather than raw cartridge storage space) is generally the major factor that limits incremental capacity of a storage system. In other words, our archive may be able to hold an additional 100TB .... but will the other associated storage resources support this much additional data?
- Our storage statistics show our workload is 1/3 writes and 2/3 reads.
- Our statistics also show that the total amount of data read by our users annually equals the amount of data currently contained in the archive (many files never read but some are read frequently!)
- Storage environment performance seems to degrade exponentially as you saturate capabilities (like an Ethernet). For this reason, we try to limit our usage to a 1/3 of theoretical data rates.

# Data Intensive Computing

- The classic problem:
  - A BIG user just finished a run on your supercomputer .... a large portion of local storage is full of his results.   How can you offload the current user's data FAST so the compute resource can be utilized?
  - Supercomputer  memory sizes, performance,  and local storage capacity has overwhelmed the capabilities of most archival storage environments.
  - Even if you have a high speed storage environment,  network speeds (and protocols)  can limit your ability to utilize the resource.
  - This problem easily overwhelms classical data caching strategy.
- Our (currently planned) strategy:   (same strategy applied to all the other computer architecture problems recently ..... )

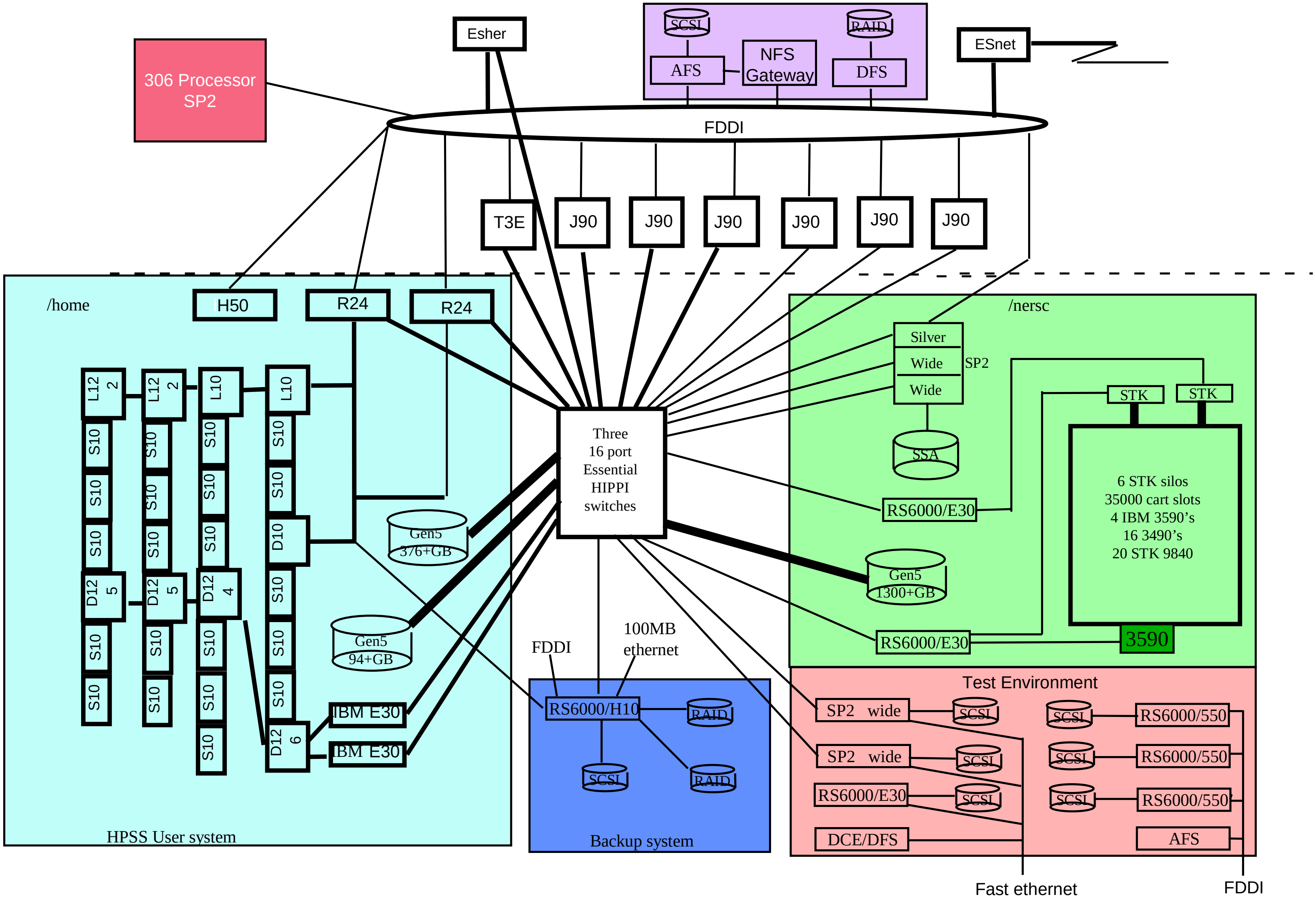
**exploit parallelism!!**

  - *STRIPE* data from supercomputer local disk directly to storage server archival devices via *MULTIPLE* network interfaces.

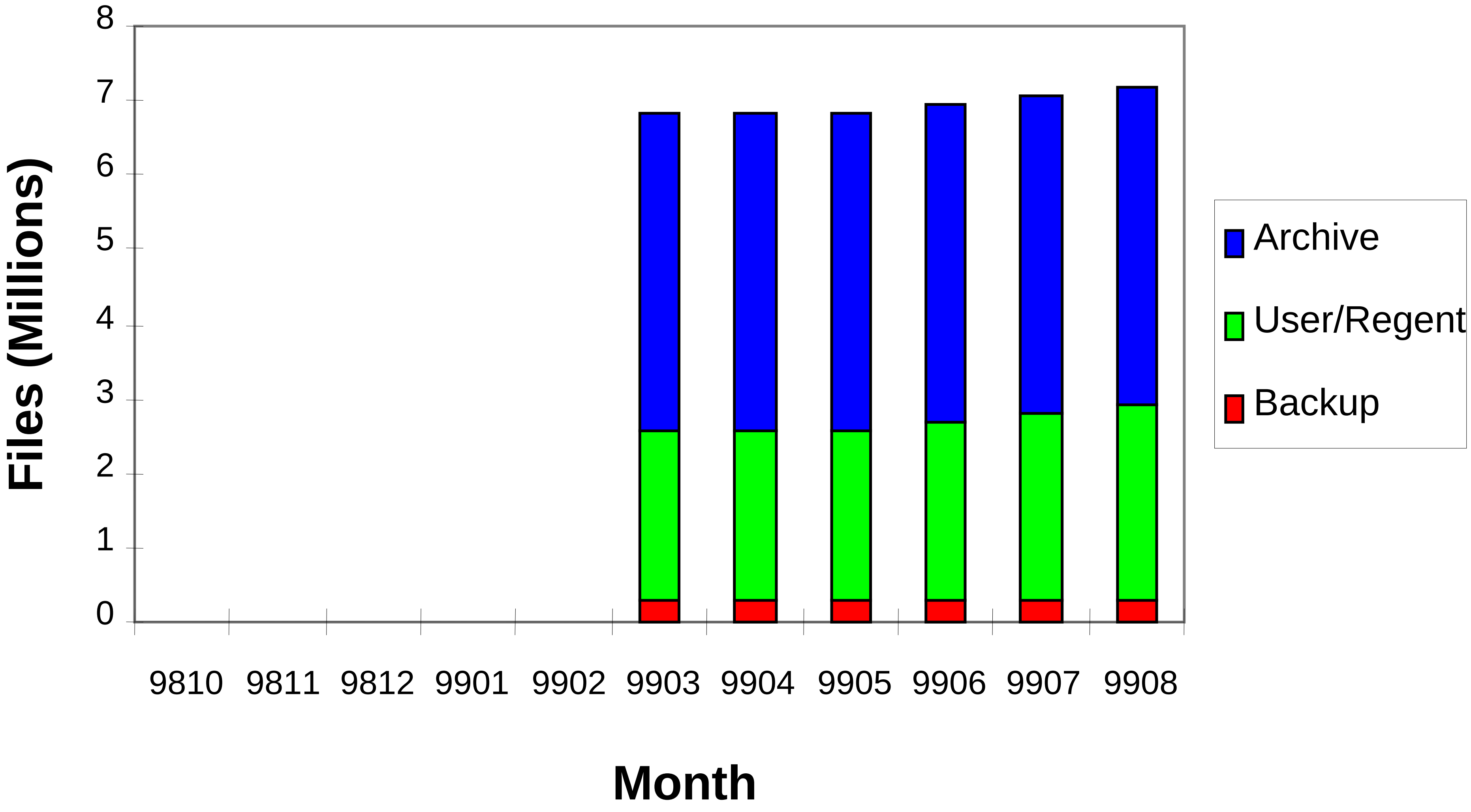
# "PROBE"
## (A DOE Sponsored HPSS Test Environment)

- Oak Ridge National Laboratory and NERSC have been funded to provide a geographically distributed HPSS test environment in which specific HPSS configurations and new devices can be tested without impacting production storage environments. ANYONE can apply for PROBE time.   We especially encourage hardware vendors to test their new high end storage products in this testbed.
- PROBE can be thought of as a "tinker-toy" environment where specific HPSS hardware and software configurations can be setup, tested, and the results documented.     Application consultants will be available to assist.
- Resources include:
  - DCE  servers  ... (AIX based at ORNL,  SUN based at NERSC)
  - HPSS Core servers platforms....  IBM currently
  - HPSS Movers .... IBM, SUN, DEC (Compact)
  - Archives .... STK silos, IBM 3494
  - Tapedrives .... IBM 3490, IBM 3590, STK 9840, STK redwood
  - Disks .... SSA, Fibre, Maxstrat, ...etc.
  - Networks ....  Gigabit Ethernet, HIPPI, 100MB Ethernet,  fibre

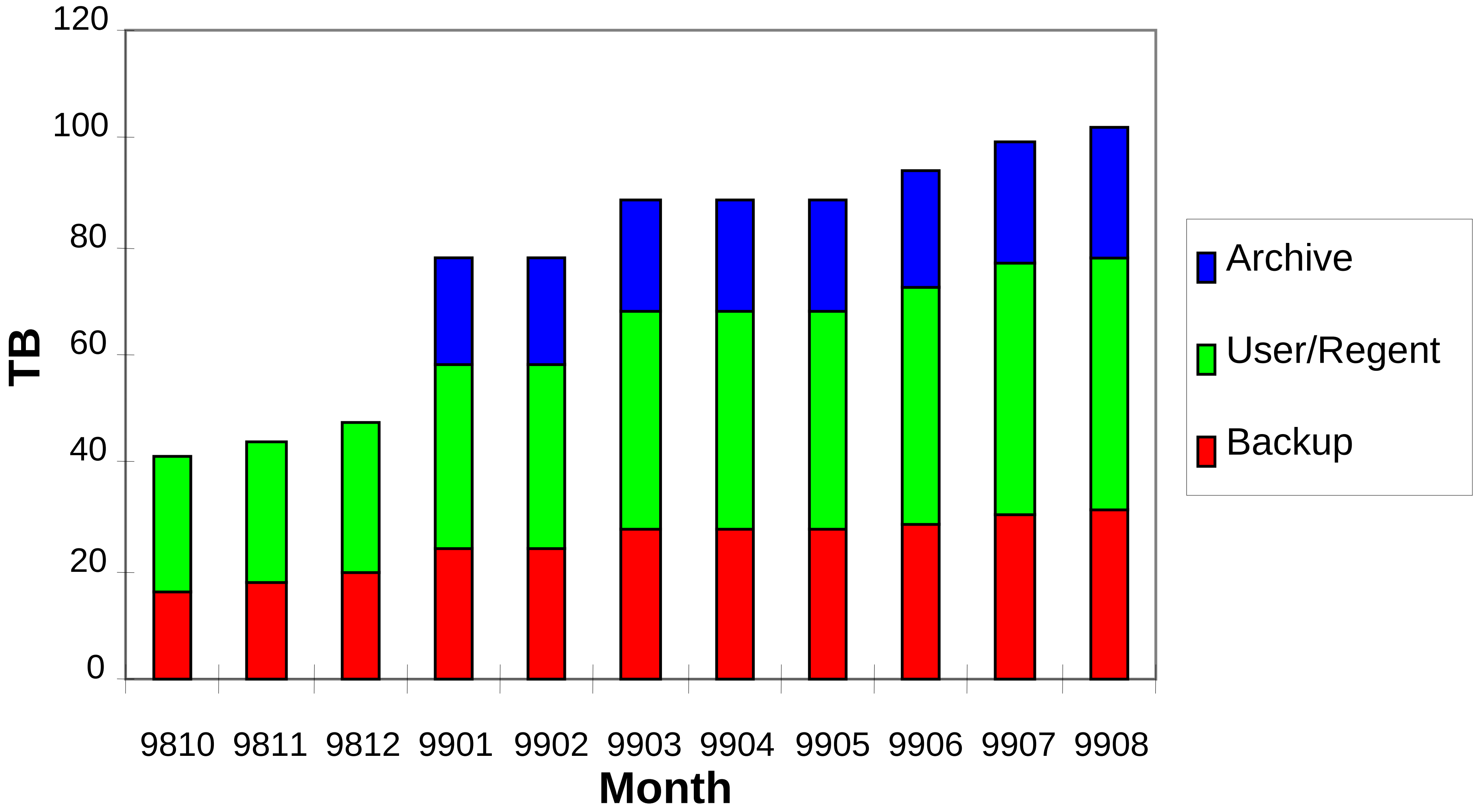
# Current NERSC Environment



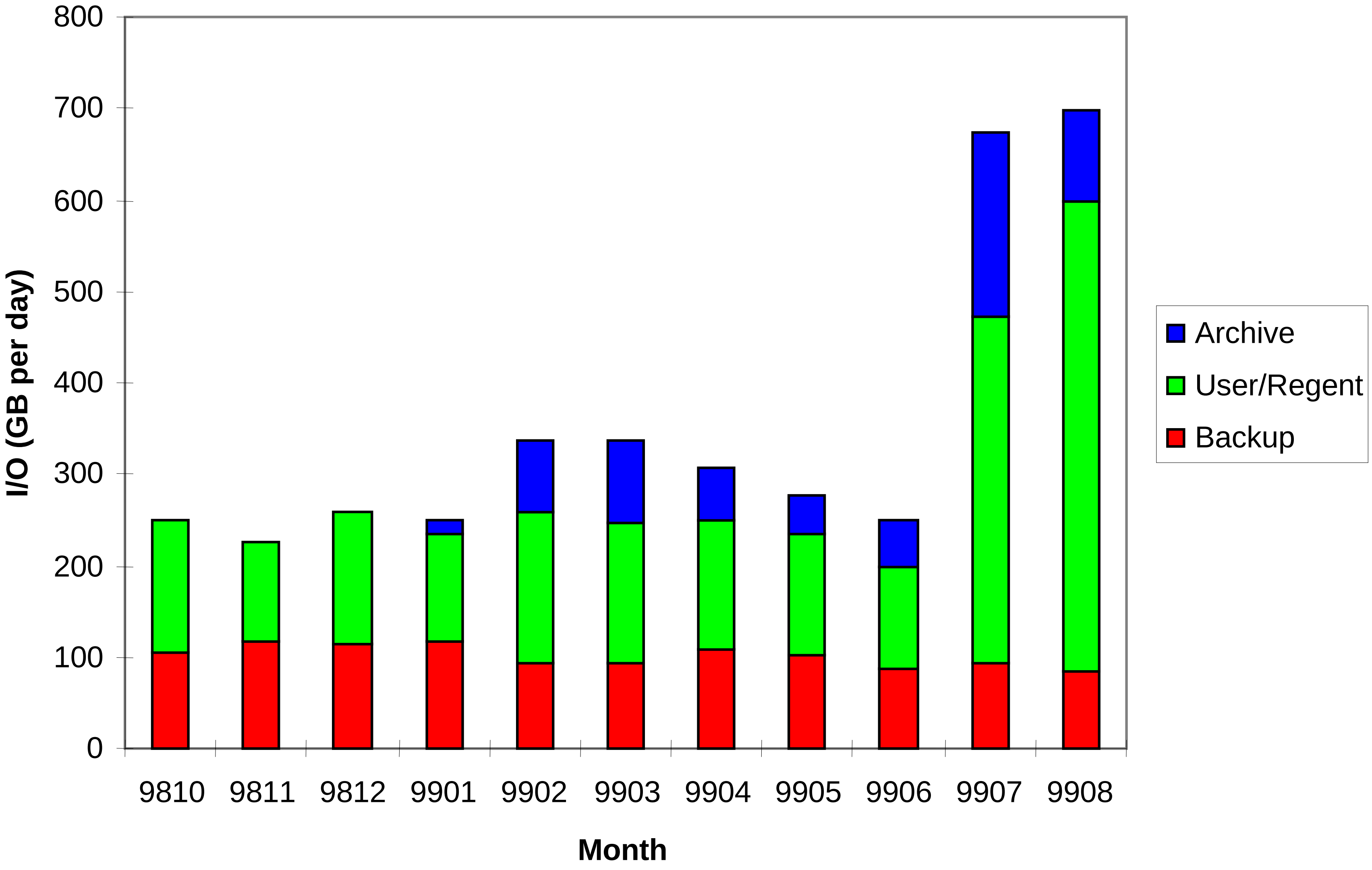
# Number of Files in Storage



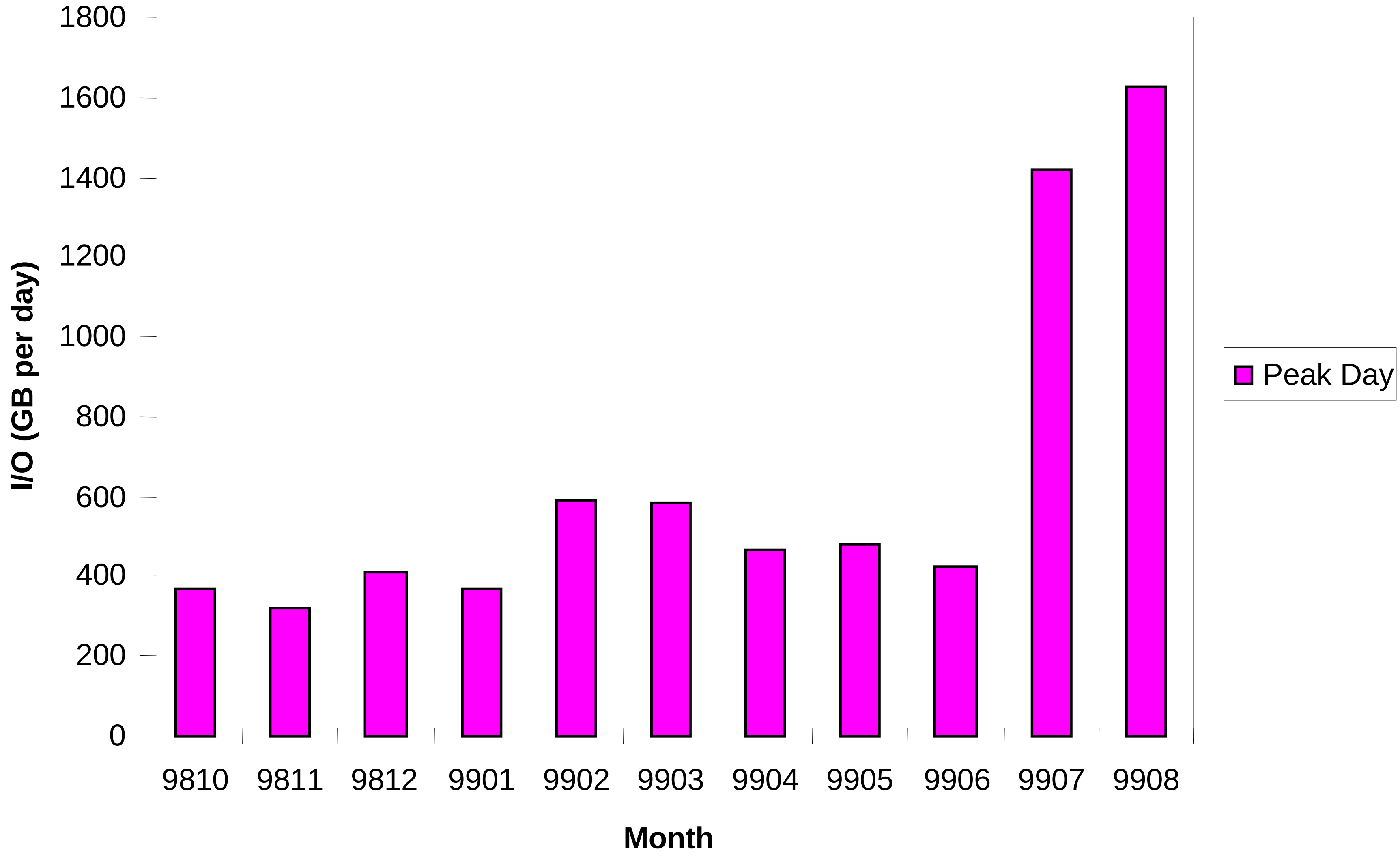
# Amount of Data Stored



# Average Daily I/O



# Peak Daily I/O



# Current System Performance

| System | Peak Rates | | | |
|---|---|---|---|---|
| | individual (MB/s) | | aggregate (MB/s @ # concurrent xfers) | |
| | write | read | write | read |
| Regent | 9 | 9 | 34@40 | 39@48 |
| Archive | 10 | 10 | 48@32 | 50@40 |

**** Tests were conducted using 200MB files (large COS)**
**Clients were two Cray J90's and two SP2 wide nodes**
**Connections were split between two HPSS ftp server nodes**
**Number of transfers/client was incremented until it peaked**

# References:

- Review of MSS systems for NERSC:

  Fitzgerald, Holmes, Hurlbert, Meyer

- A back of the Envelope Look at File Storage Bandwidths

  Fitzgerald, Holmes, Hurlbert, Meyer

- Plans for Archivel Storage Quotas

  Holmes, Fitzgerald, Daveler, Hurlbert, Meyer